View metadata, citation and similar papers at core.ac.uk

brought to you by CORE

provided by AIS Electronic Library (AISeL)

Association for Information Systems
AIS Electronic Library (AISeL)

PACIS 2006 Proceedings

Pacific Asia Conference on Information Systems (PACIS)

2006

Applying Kohonen Vector Quantization Networks for Profiling Customers of Mobile Telecommunication Services

Indranil Bose
The University of Hong Kong, bose@business.hku.hk

Chen Xi
The University of Hong Kong

Follow this and additional works at: http://aisel.aisnet.org/pacis2006

Recommended Citation
Bose, Indranil and Xi, Chen, "Applying Kohonen Vector Quantization Networks for Profiling Customers of Mobile Telecommunication Services" (2006). *PACIS 2006 Proceedings*. 61.
http://aisel.aisnet.org/pacis2006/61

This material is brought to you by the Pacific Asia Conference on Information Systems (PACIS) at AIS Electronic Library (AISeL). It has been accepted for inclusion in PACIS 2006 Proceedings by an authorized administrator of AIS Electronic Library (AISeL). For more information, please contact elibrary@aisnet.org.

Applying Kohonen Vector Quantization Networks for Profiling Customers of Mobile Telecommunication Services

Indranil Bose
School of Business
The University of Hong Kong

Chen Xi
School of Business
The University of Hong Kong
bose@business.hku.hk

Abstract

Customer clustering is used to understand customers' preferences and behaviors by examining the differences and similarities between customers. Kohonen vector quantization clustering technology is used in this research and is compared with K-means clustering. The data set consists of customer records obtained from a mobile telecommunications service provider. The customers are clustered using various attributes that are broadly grouped under usage, revenue, handset, and service. The clustering results are examined to see the relationships between different types of attributes. The analysis leads to the discovery of several interesting facts about customers that may be of use to mobile service providers.

Keywords: Clustering, Customer profiling, Kohonen Vector Quantization, Mobile telecommunication

1. Introduction

Due to the fast penetration of mobile phones, mobile telecommunication services are becoming more and more popular. In Hong Kong, the penetration rate of mobile phones is more than 100% which means on an average people own more than one mobile phone. On the other hand, there are 7 mobile telecommunication operators in Hong Kong. So the competition to acquire and retain customers among mobile service providers is fierce. The key to survival in this competitive industry lies in knowing the customers better. Different people have different preferences for using mobile telecommunication services and mobile phones. According to IDC group's study of usage patterns of mobile data services across the Asia Pacific region SMS is the most popular mobile service used (IDC, 2006). About 65% of the customers send SMS everyday. Only 35% of customers do not use SMS that frequently. Treating all customers without differentiation may lead to the situation that some customers have to choose services they do not want and this may lead to loss of customers. One of the approaches used to understand customers is customer clustering. Clustering classifies customers into different groups in order to see similarities and dissimilarities between customers. After clustering customers, mobile service providers can develop different mobile telecommunication services for different clusters in order to match the services to the customers' preferences. For example, a plan with

large number of free roaming minutes can be offered to people who use roaming services a lot. Usually, clustering of customers make use of demographic or customer behavior data. In mobile marketing, the usage information of mobile telecommunication services is the best data that can be used to reflect customers' behaviors and preferences. In this research, we use mobile telecommunications usage data to cluster customers and discover the patterns in their behavior.

This paper starts with the discussion of the related literature and moves on to a description of the research methodology. The third and fourth parts describe the experimental results and related discussion on the results respectively. In the final section possible future directions of research are elaborated.

2. Literature Review

Customer clustering aims to classify customers into different groups. Customers within the same group are said to be more similar to each other than to customers in different groups. Customers are clustered according to their characteristics. Typically, there are two kinds of data on customers' characteristics that are used in mathematical models for direct marketing. One type of data includes customers' geographic, demographic, lifestyle, and socio-graphic characteristics (Bult, 1993; Bult and Wittnk, 1996). Demographic data includes information on customers' age, sex, and family size etc. Lifestyle data includes information on customers' habits, booking of a certain magazine, leisure interests etc. Geographic data includes information on customers' location of home, office, and business. Another kind of data that is used concerns customers' interactive behavior with marketers. Data on customers' behavior includes customers' transaction records, feedback from customers and web browsing records. It is believed that customers' transaction records can reveal valuable information on customers' behavioral patterns. RFM (Roberts and Berger, 1989) variables are the most prevalently used behavioral data in marketing research where R stands for 'recency', which measures the length of time since a customer's last buying activity or the number of consecutive solicitations without response from the customers after the last purchase. F stands for 'frequency', which measures the number of products bought by a customer during a period. M stands for 'monetary', which is the monetary value of a customer's spending during a certain period or in the last buying transaction. Due to the ubiquity of Internet access and e-commerce, more and more customers search for product information and even buy products online. Thus, data on customers' browsing and purchasing activities are also used to analyze customers' behavior (Suh et al., 2004; Yu et al., 2005)

Various approaches have been used for clustering. One type of approach is to select a certain cutoff for independent variables so that the segmentation can have highest profit gain. One example of this approach is RFM clustering. Jonker et al. (2004) and Bitran and Mondschein (1996) classified customers using RFM values. Customers are first divided into groups based on their R value. Within each group, customers are grouped according to their F value and then grouped again based on their M value. The second type of clustering approach applies statistical models such as latent class analysis. Latent class analysis captures heterogeneity among customers by first specifying the number of clusters and then running regression models to estimate the value of coefficients. Gonul et al. (2000), Desarbo and Ramaswamy (1994), and Wedel et al. (1993) used latent class analysis to model the customer response problem. The third type of clustering approach

involves the use of data mining techniques such as K-means clustering and Self-Organizing Map (SOM). K-means and SOM are examples of unsupervised machine learning techniques. The K-means technique classifies customers into a specified number of clusters while SOM decides the number of clusters automatically. SOM consists of a set of artificial neurons whose weights are adjusted to match input vectors in a training set (Kohonen, 1995). Min and Han (2005) used SOM to cluster customers with similar interests at different periods of time. Weng and Liu (2004) used a two-stage clustering approach which integrated SOM and K-means clustering analysis. From the results of their experiments, it is known that two stage clustering has higher density of clusters than single SOM clustering. However, mobile telecommunication data is different from the data used in this type of research related to direct marketing. The data used in direct marketing research usually includes catalog selling records or charity donation information. These types of transactions occur once or twice a year. In contrast to that, mobile telecommunication users may use the mobile telecommunication services anytime and anywhere within a single day. Hence the frequency of mobile communication usage data is higher.

In research related to mobile telecommunication industry, the most popular topic is detection of churning. Churning means customers switching their accounts from one mobile telecommunication provider to another. Bhattacharyya (1999), Bhattacharyya (2000), and Wei and Chiu (2002) studied churning detection problems using mobile telecommunication usage data. However, clustering is not indicated as a possible technique in any of these papers. In practice, mobile telecommunication usage data contains hundreds of attributes and therefore an important task is to find most appropriate attributes as independent variables to predict churning. In this paper, we cluster customers using different data about customers and compare the relationship between different clustering approaches.

3. Research Methodology

Data mining techniques are employed in this research and the data mining process model of CRoss Industry Standard Process, CRISP-DM in short, is followed (CRISP, 1999). This process model describes a standard data mining project lifecycle (Wirth and Hipp, 2000; Gersten et al., 2000). The CRISP-DM process has 6 phases: business understanding, data understanding, data preparation, modeling, evaluation and deployment, as shown in Figure 1 (adapted from CRISP, 1999). We have stated our business understanding in the introduction and since this is research in progress we do not discuss the deployment phase.

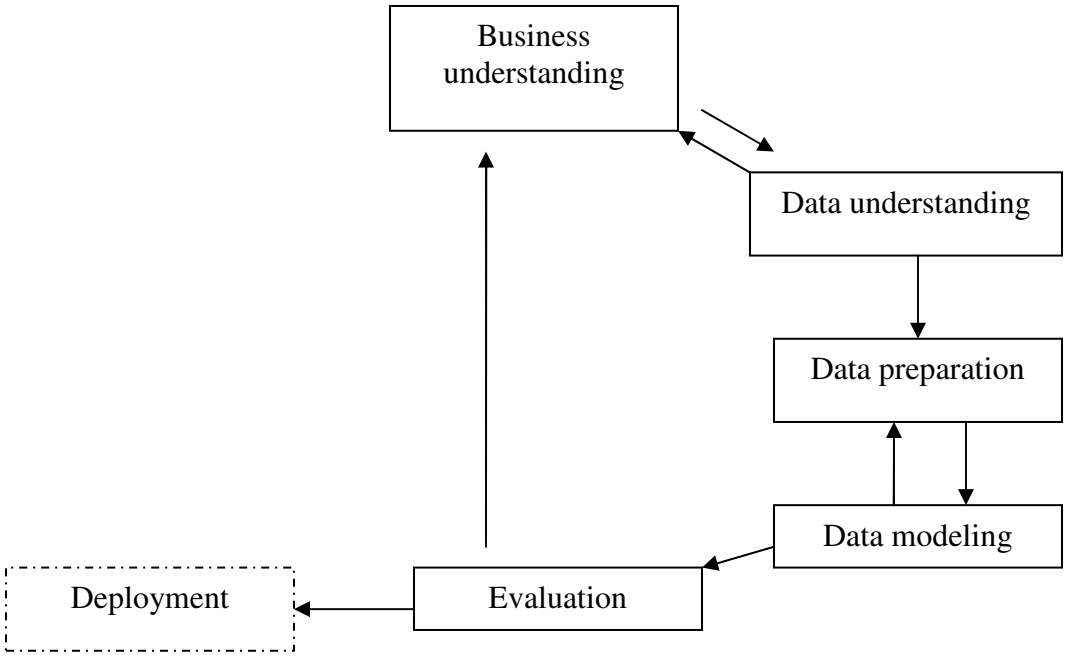

Figure 1: CRISP-DM (adapted from CRISP, 1999)

3.1 Data Collection and Understanding

For this study, we collected mobile telecommunication data from a major local mobile telecommunication operator. There are 52374 records of customers. Each of these records represents a single mobile telecommunication user and is characterized by 223 attributes. The data can be grouped into several types as shown in Table 1. The usage information is recorded in terms of minutes of usage (MOU). For each type of usage information, MOU for each of last three months and aggregation of MOU over last six months, last nine months, and last twelve months are included. For each type of revenue information, there are information on revenue of each of the last three months and aggregated revenue of last three months, last six months, last nine months, and last twelve month. Usage, revenue, age, and tenure are continuous variables while other variables are discrete variables.

Type of attribute	Description
Account information	creation date, payment method
Customer information	age, gender, customer type, communication approach information
Subscription information	plan group, tenure
Handset features	handset brand, support features
Handset purchase information	purchase date, purchased in last 12 months, purchased in last 3 months
Usage information	mobile IDD, roaming, outbound roaming, GPRS, PHS, PM, SMS
Services	voice service, data service, IDD& roaming service, other services
Revenue information	net revenue, VAS revenue, IDD revenue, outbound roaming revenue

Table 1: Attributes for mobile service customers

3.2 Data Cleaning and Preprocessing

Attributes which have more than 60% missing data and attributes that have only one value are removed as the first step of data cleaning. It is found that most of the continuous type variables are highly skewed in their distribution. 135 out of 160 continuous type variables have skewness index higher than 2. The average absolute value of skewness is 10.67 and the range varies from 0.64 to 81.1. To overcome the problem of skewness, we transformed these variables using the log function. After transformation, the average absolute value of skewness reduced to 1.29 and the range of absolute value of skewness varied from 0.02 to 11.6. This data set has discrete variables with many values. For example, handset model has 39 values, customer type has 9 values, and poly ringtone has 13 values. Since too many values of discrete variables will increase the calculation burden therefore we transform these variables. We re-group customer type according to whether they are personal user or not. For the attribute of hand set model, we keep famous brands with high frequencies such as Nokia, Sony Ericsson, Sharp, and Motorola and combine other brands into two groups: other Japan/Korea brand and remaining brands.

3.3 Data Modeling

The main purpose of data modeling is to discover interesting information from this data set. Marketing or research opportunities come from good understanding of customers' behaviors and preferences. Customer clustering is one of the most commonly used approaches to understand customers because it can group customers together and show similarities and dissimilarities between them. Since we were exploring the data set and did not have any target, unsupervised clustering techniques were used. We used K-means clustering and Kohonen Vector Quantization networks (KVQ). Kohonen Vector Quantization networks are unsupervised clustering techniques closely related to K-means cluster analysis and SOM (Kohonen, 1995). K-means begins clustering by selecting K (specified number of clusters) seeds (centers of clusters) according to the distribution of the data set. And these seeds selected by K-means tend to be approximately uniformly distributed. In other words, K-means assumes a uniform distribution of the data set. In contrast, KVQ selects code book vectors, which is KVQ's version of seeds, randomly and the distribution probability density function is approximated by a set of optimally placed discrete parameter vectors. Code book vectors which are closest to each cluster are found and moved closer to the clusters by a certain portion. The portion is specified by the learning ratio. KVQ and SOM have similar learning algorithms but SOM considers both distances in input space and distances in the map while KVQ only considers the former. The results of two clustering techniques are compared. The criteria used for comparison are cluster numbers, average distances, average CV indexes (ratio of standard deviation and mean of distances). Distances in the context of clustering refer to Euclidean distances between data points and the seeds of clusters. Lower average distance or lower CV index means higher cluster concentration.

This data set has more than 200 attributes. Using all of them in clustering is not reasonable because of the difficulty in interpreting clustering results. Thus we have to make choices among the 200 attributes. The first type of attributes we choose are the

usage attributes because they reflect customers' behaviors directly. The second type of attributes we choose are the revenue attributes because they reflect company's profitability. The third type of attributes we choose are the subscription information and the information on services registered. In mobile telecommunication, these services and plans are the products customers buy from mobile telecommunication operators and therefore they reflect the users' preferences. The last group of information chosen is about user's handset. The functions of mobile phones may influence the usage patterns of customers and may also reflect the users' preferences. By clustering customers based on these four groups of attributes we create four profiles of customers: usage profile, revenue profile, service profile, and handset profile.

As the first step, we use K-means clustering and KVQ to cluster the data set using four groups of attributes. The attributes used in handset group, service group, and revenue group are shown in Table 2, Table 3, and Table 4 respectively. The attributes in usage group are usage of first three months and total usage of last six month, last nine month and last twelve months of the following services: mobile IDD, roaming, outbound incoming roaming, outbound outgoing roaming, PHS, PM call, SMS, and GPRS call. Since the list is quite large we do not show them in a table.

Name	Measurement	Description
MOBILE_GAME	binary	support or not
MMS	binary	support or not
LOGO	binary	support or not
PICTURE	nominal	type of picture
WAP_PUSH	binary	support or not
WAP	binary	support or not
THREEG	binary	support or not
RINGTONG	nominal	type of ringtone
GPRS	binary	support or not
HANDPHONE	nominal	brand

Table 2: Handset attributes

Name	Measurement	Description
DATASET_2_GPRS	binary	subscribed or not
CONNECTING_TONE	binary	subscribed or not
IDD	binary	subscribed or not
SMC_STEALTH_ROAMING	binary	subscribed or not
SMC_ICF	binary	subscribed or not
ROAMING	binary	subscribed or not
SMS_VIA_EMAIL	binary	subscribed or not
PICTUREMAIL_REGISTRATION	binary	subscribed or not
DATA___WAP_SERVICE	binary	subscribed or not
TENU_MRH	interval	log of tenure

Table 3: Service attributes

Name	Measurement	Description
IDD__QIO	interval	log of last 6 month total IDD revenue
NET__WKH	interval	log of last 6 month total net revenue
OR_R_EQ4	interval	log of last 6 month total outbound roaming revenue
OTHE_77I	interval	log of last 6 month total other revenue
VAS__LP1	interval	log of last 6 month total VAS Non voice net revenue
VAS__IU4	interval	log of last 6 month total VAS voice net revenue

Table 4: Revenue attributes

3.4 Results and Evaluation

In Table 5, for the row titled number of clusters, the first number represents the number of clusters using KVQ and the next number represents the number of clusters using K-means. K-means clustering using service attributes failed to identify appropriate number of clusters. K-means clustering using handset attributes identified 2 clusters, one of which contains about 80% of the whole population. In contrast, KVQ clustering produces more evenly distributed clusters.

In terms of CV index, for the clustering using usage attributes, the performances of the two techniques are similar. For the clustering using attributes of revenue group, K-means has higher CV index than KVQ. For clustering using handset attributes and service attributes, the K-means approach has lower CV index. However, in terms of average distance, KVQ always performs better than K-means.

	usage	revenue	handset	service
KVQ	0.2414	0.4310	0.2424	0.5366
K-means	0.2383	0.5238	0.1382	0.4190
Number of clusters	6 and 6	5 and 5	5 and 2	4 and N/A

Table 5: Comparison of CV index and number of clusters between K-means and KVQ

	usage	revenue	handset	service
KVQ	5.4816	1.9988	1.3623	2.8554
Kmeans	7.0108	2.3373	1.4992	2.9484

Table 6: Comparison of average distance between KVQ and K-means

From the results we can see that KVQ is more capable of identifying differences between customers and clustering them into evenly distributed groups. Thus, we use the results of KVQ clustering for further analysis. We begin with clustering using usage attributes. 67 usage attributes are used to cluster the customers. All of them are standardized and log transformed. The KVQ approach rates the importance of these variables in the formation of clusters. The rating varies from 0 to 1 where 1 identifies the most important attribute. Since it is hard to interpret the clusters using all 67 attributes, we choose only 16 of them which have importance rate greater than 0. These attributes are shown in Figure 2. In

Figure 2, from left to right, these attributes are GPRS usage information, mobile IDD usage, outbound roaming usage, PM usage, and SMS usage, grouped according to the facts they reflect. Group five has highest GPRS usage. Group four has highest mobile IDD, outbound roaming and roaming usage. Group 3 has highest SMS usage. For each type of mobile service, we can identify from Figure 2, three relative levels of usage: high, medium, and low. Thus, we can label them in the way as shown in Table 7.

Clustering customers according to revenue attributes uses 7 revenue attributes, as shown in Figure 3. We show the attributes in Table 8. Table 9 and Table 10 show the results of clustering using handset attributes and service attributes respectively. The value in each cell for the discrete variables represents the percentage of people who registered corresponding services or handsets which support the corresponding features.

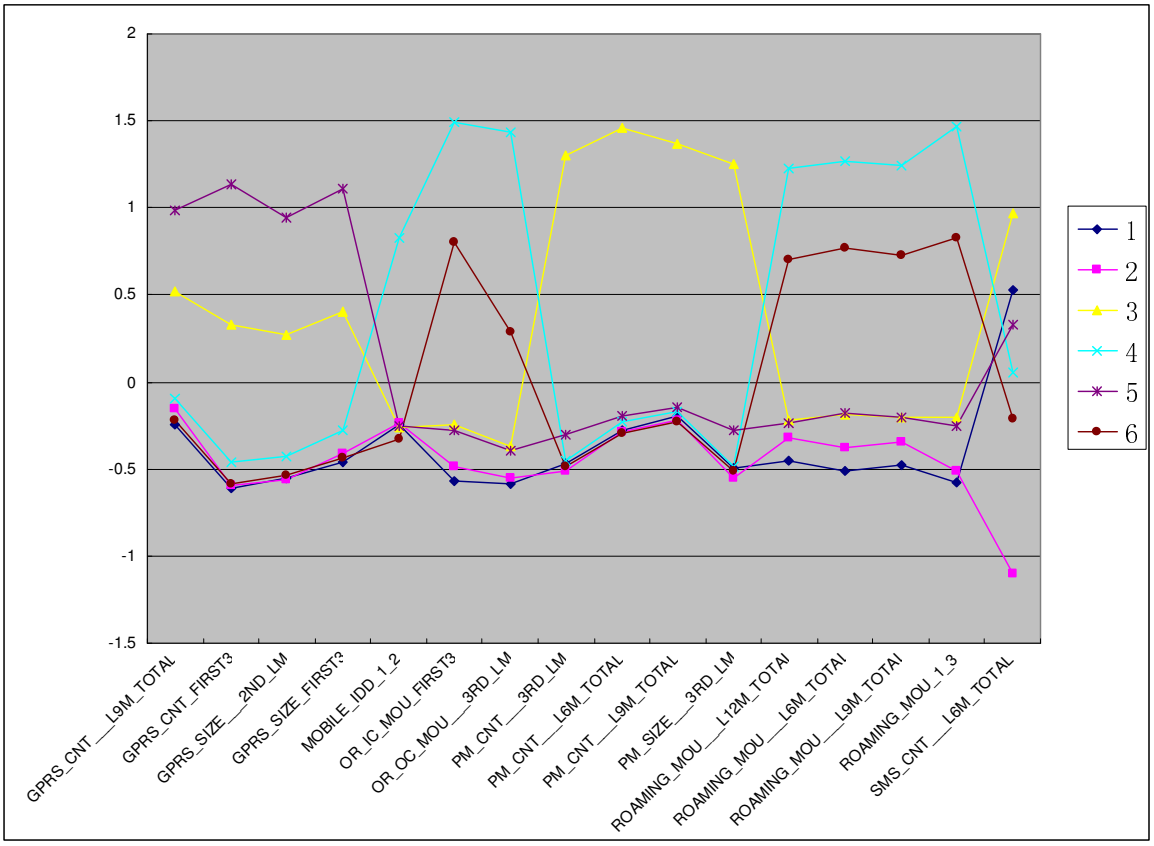

Figure 2: Usage based clusters

Cluster	GPRS	Mobile IDD	Outbound roaming	PM call	Roaming	SMS
1	low	low	low	low	low	medium
2	low	low	low	low	low	low
3	medium	low	low	high	low	high
4	low	medium	high	low	high	medium
5	high	low	low	low	low	medium
6	low	low	medium	low	medium	medium

Table 7: Usage based profiles

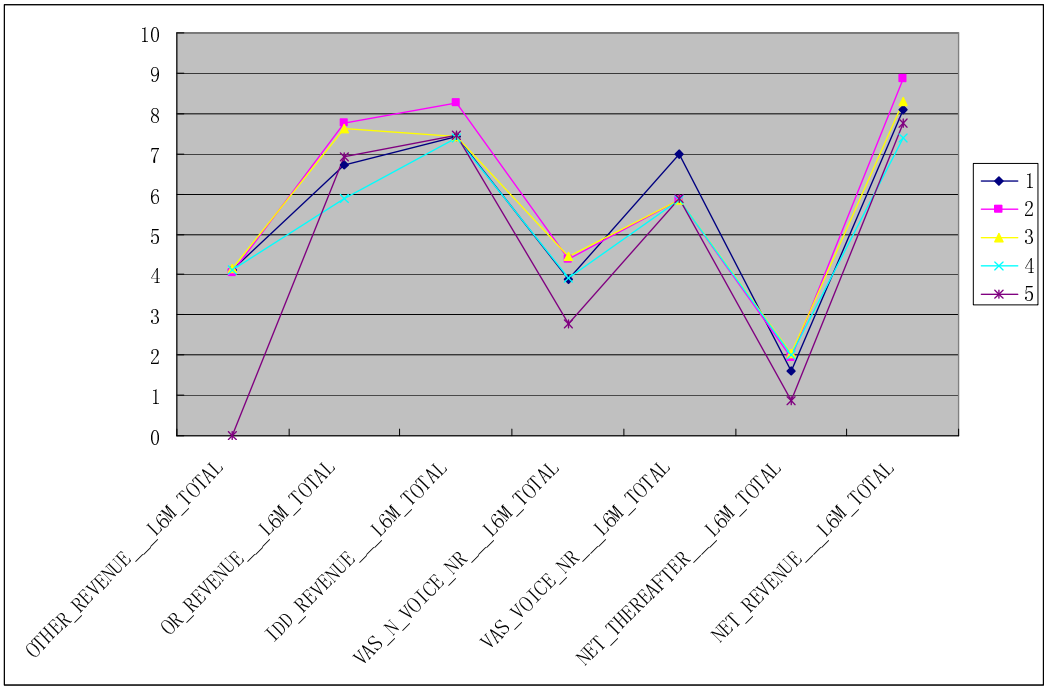

Figure 3: Revenue based clusters

OTHER_ REVENUE _L6M_ TOTAL	OR_ REVENUE_ L6M_ TOTAL	IDD_ REVENUE_ L6M_ TOTAL	VAS_N_ VOICE_NR_ L6M_ TOTAL	VAS_ VOICE_ NR_ L6M_ TOTAL	NET_ THERE AFTER_ L6M_ TOTAL	NET_ REVENUE_ L6M_ TOTAL
high	medium	low	medium	high	medium	medium
high	high	high	high	low	high	high
high	high	low	high	low	high	medium
high	low	low	medium	low	high	low
low	medium	low	low	low	low	medium

Table 8: Revenue based profiles

	1	2	3	4	5
Mobile_game=S	0.022	0.997	0.003	1.000	0.982
Mms=S	0.426	0.998	0.007	1.000	1.000
Logo=S	0.381	0.011	0.287	0.998	0.000
Picture=EMS PICTURE	0.153	0.891	0.000	0.003	0.000
Picture=NO	0.411	0.000	0.718	0.000	0.935
Picture=NOKIA AND EMS	0.001	0.003	0.000	0.001	0.045
Picture=NOKIA PICTURE	0.435	0.021	0.282	0.997	0.002
Wap_push=S	0.851	0.998	0.000	1.000	0.914
Wap=S	1.000	1.000	0.051	1.000	1.000

ThreeG=S	0.023	0.012	0.000	0.002	0.343
Ringtong=high	0.000	0.578	0.007	0.076	0.507
Ringtong=low	0.970	0.104	0.979	0.566	0.223
Ringtong=mid	0.030	0.318	0.014	0.359	0.097
ringtong=sup	0.000	0.000	0.000	0.000	0.174
GPRS=Y	0.978	0.979	0.979	0.981	0.978

Table 9: Handset profiles

Services	1	2	3	4
log(TENURE)	6.518	6.930	7.024	7.076
Dataset_2_GPRS=S	0.798	0.846	0.859	0.866
Connecting_Tone=Y	0.254	0.222	0.194	0.159
IDD=Y	0.172	0.646	0.627	0.838
SMC_Stealth_Roaming=Y	0.000	0.067	0.051	0.119
SMC_ICF=Y	0.007	0.048	0.050	0.168
Roaming=Y	0.000	1.000	1.000	0.991
SMS_via_Email=Y	0.000	0.000	0.000	1.000
PictureMail_Registration=Y	0.664	0.718	0.727	0.741
Data___WAP_Service=Y	0.881	1.000	0.000	0.942
Top 3 most used plans	super, high, basic	high, medium, basic	medium, high, basic	special, high, medium

Table 10: Service profiles

For usage clustering and revenue clustering, we count the number of 'high', 'medium' and 'low' for each cluster and rank them accordingly. The higher the ranking the lower the usage or revenue respectively. For service clustering, we add up all the cells to calculate a score for each cluster. Higher score means that customers in that cluster registered more service than other customers in other groups. For handset clustering, we use a similar approach as the one for service clustering. The difference is that we assign weights to 'ringtone' attribute and 'picture' attribute. 'Ringtone=sup' is given a weight of 3. 'Ringtone=high' is given a weight of 2, 'Ringtone=mid' is given a weight of 1. 'Ringtone=low' is given a weight of 0. 'Picture=nokia and ems' is given a weight of 2. 'Picture=nokia' and 'picture=ems' are given weights of 1. 'Picture=no' is given a weight of 0. Higher score means that the people within this cluster use more advanced handset models.

Clusters	Usage	Revenue	Service	Handset
1	5	3	9.2938	4.3013
2	6	1	11.4770	7.3868
3	2	2	10.5322	1.6354
4	1	4	12.8989	7.4919
5	3	5		6.9422

6	4			

Table 11: Rankings of clusters

In Table 12, we examine the relationship between service clusters and revenue clusters. The value in each cell after '\' represents the percentage of customers who belong to the cluster while the value in each cell before '\' represents the frequency. From the previous analysis, we know that revenue clusters 2, 3, and 1 are the three clusters which have the three highest revenues while cluster 5 has the lowest revenue. All these clusters have high percentage of people belonging to service cluster 2, the group with 2nd highest service score.

service\revenue	1	2	3	4	5
1	655\0.03	139\0.01	13\0.01	31\0.02	12\0.01
2	24429\0.94	17295\0.96	1517\0.94	1509\0.90	1359\0.96
3	817\0.03	549\0.03	58\0.04	50\0.03	28\0.02
4	78\0.00	93\0.01	24\0.01	78\0.05	11\0.01

Table 12: Relationship between service and revenue clusters

In Table 13, relationships between usage and revenue clusters are shown. For revenue cluster 2, 43% belongs to usage cluster 6 which ranks 4th among usage clusters. 43% of revenue cluster 1 customers belong to usage cluster 1 while revenue cluster 1 ranks 3rd among revenue clusters and usage cluster 1 ranks 5th among usage clusters. Many members of usage cluster 4 (which ranks 1st among usage clusters) belong to revenue cluster 2. This indicates that high usage contributes to high revenue. Again, the second and third usage cluster, cluster 3 and cluster 5 also have large portions of customers belonging to revenue cluster 2. Except usage cluster 4, other usage clusters have more customers belonging to revenue cluster 1 than belonging to revenue clusters 2 and 3.

usage\revenue	1	2	3	4	5
1	11135\0.43	2521\0.14	402\0.25	442\0.26	283\0.20
2	7757\0.30	1267\0.07	432\0.27	452\0.27	229\0.16
3	1932\0.07	1816\0.10	54\0.03	29\0.02	62\0.04
4	35\0.00	2037\0.11	165\0.10	218\0.13	693\0.49
5	3289\0.13	2688\0.15	160\0.10	61\0.04	75\0.05
6	1831\0.07	7747\0.43	399\0.25	466\0.28	68\0.05

Table 13: Relationship between usage and revenue clusters

Table 14 shows relationships between handset clusters and revenue clusters. In revenue cluster 2, most of the customers belong to handset clusters 2, 4, and 5. The distribution of customers in revenue clusters 1 and 3 are similar to that of cluster 2. Handset clusters 2, 4, and 5 are clusters whose customers own more advanced handsets than customers in other

clusters. It is interesting to note that although customers in handset cluster 5 do not have advanced handsets, they contribute more customers to revenue clusters 1 and 2.

handset\revenue	1	2	3	4	5
1	2309\0.09	1309\0.07	154\0.10	214\0.13	147\0.10
2	5006\0.19	3601\0.20	304\0.19	315\0.19	265\0.19
3	4078\0.16	2199\0.12	217\0.13	371\0.22	227\0.16
4	7264\0.28	5323\0.29	536\0.33	513\0.31	484\0.34
5	7322\0.28	5644\0.31	401\0.25	255\0.15	287\0.20

Table 14: Relationship between handset and revenue clusters

4. Discussion of Results

Services registered and revenues generated by customers are not closely related. Most people register available services but do not use them often because most services charge customers money only when they use them. This may indicate that either the customers think the usefulness of the mobile services provided to them are low or they have not realized the usefulness of these services. Thus, researchers and marketers of mobile telecommunications can think of studying the customers' preferences and develop more personalized mobile communication services for their customers. Secondly, customers' usage of mobile communication services and revenues they generate are unbalanced. Some customers who have low usage of services contribute high revenues. This may be good for the operators at present but may be dangerous in future. When these customers realize the unbalance between the money they pay and the service they use, they may either switch to plans with low revenue or even switch to other companies. Marketers should take suitable actions to either provide these customers more valuable services or advise them to use more suitable plans with lower fees. For researchers, when studying customer churning problems, the factors related to unbalance between usage and revenue should be taken into consideration. We also notice that customers that have high usage contribute less money and this phenomenon may disclose customers' potential to generate more profit. Packages with higher level of services may be promoted to these customers. As a result, if researchers want to include the factor of handset in models studying customers' mobile telecommunication behavior patterns, it may not be necessary to develop many levels of value for that factor.

From Figure 2, one interesting phenomenon can be identified. Cluster 6 and cluster 2 have very similar pattern in terms of GPRS usage, PM usage, mobile IDD usage, outbound roaming, and roaming usage. However they behave very differently in terms of SMS usage. For marketers, there may be opportunities to make the customers in cluster 2 use more SMS. For researchers, it may be interesting to compare these two clusters to see the reason of the discrepancy in SMS usage pattern.

5. Future Work and Conclusion

In this study, we explored the data set of mobile telecommunication users. We performed customer clustering using K-means and KVQ and compared their performances. In this case, KVQ performed better than K-means in terms of CV index and average distance. However, the way we compared two methods is relatively simple. Thus, it may not be

appropriate the claim the superiority of KVQ over k-means yet. More sophisticated clustering validation methods should be applied to compare the two methods to reach a solid conclusion. We examined the relationships between different clustering approaches. By finding discrepancies between clustering approaches, some marketing or research opportunities are identified. The discrepancy between usage clusters and revenue clusters is worth further study. Overall higher usage does not indicate high revenue which means the existence of marketing opportunities. On the other hand, it implies that other factors are influencing revenue or some service usages have more influence on revenue than others. As the next step, we can cluster customers within the usage clusters using handset and service information. We can also cluster customers by usage information of each type of service in order to see the influence of that service on revenue.

6. References

Bhattacharyya, S. "Direct Marketing Performance Modeling Using Generic Algorithms," *INFORMS Journal on Computing* (11:3), 1999, pp. 248-257.

Bhattacharyya, S. "Evolutionary Algorithms in Data Mining: Multi-objective Performance Modeling for Direct Marketing," *Proceedings of the Sixth International Conference on Knowledge Discovery and Data Mining,* Boston, MA, USA, 2000, pp. 465-473.

Bitran, G. R., and Mondschein, S. V. "Mailing Decisions in the Catalog Sales Industry," *Management Science* (42:9), 1996, pp. 1364-1381.

Bult, J. R. "Semiparametric versus Parametric Classification Models: an Application to Direct Marketing," *Journal of Marketing Research* (30:3), 1993, pp. 380 -390.

Bult, J. R., and Wittnk, D. R. "Estimating and Validating Asymmetric Heterogeneous Loss Functions Applied to Health Care Fund Raising," *International Journal of Research in Marketing* (23), 1996, pp. 215 – 226.

CRISP-DM. *The CRISP-DM Process Model for Data Mining*, http://www.crisp-dm.org, 1999.

Desarbo, W. S., and Ramaswamy, W. "CRISP: Customer Response-based Interative Segmentation Procedures for Response Modeling in Direct Marketing," *Technical Working Paper*, Marketing Science Institute, Cambridge, Massachusetts, 1994.

Gersten, G., Wirth, R., and Arndt, D. "Predictive Modeling in Automotive Direct Marketing: Tools, Experiences and Open Issues," *Proceedings of the Sixth International Conference on Knowledge Discovery and Data Mining,* Boston, MA, USA, 2000, pp.398-406.

Gonul, F., Kim B. D., and Shi, M. Z. "Mailing Smarter to Catalog Customers," *Journal of Interactive Marketing* (14:2), 2000, pp. 2-16.

IDC**,** *IDC Survey Indicates That Less Than 10% of Users are Utilizing Services other than SMS*, http://www.idc.com/getdoc.jsp?containerId=pr2006_03_03_130022, 2006.

Jonker, J-J., Piersma, N., and Van den Poel, D. "Joint Optimization of Customer Segmentation and Marketing Policy to Maximize Long-term Profitability," *Expert Systems with Applications* (27), 2004, pp. 159-168.

Kohonen, T. *Self-organizing maps,* Springer, Berlin, Germany, 1995.

Min, S-H., and Han, I. "Detection of the Customer Time-variant Pattern for Improving Recommender Systems," *Expert Systems with Applications* (28), 2005, pp. 189-199.

Roberts, M. L., and Berger, P. D. *Direct Marketing Management*, Prentice-Hall, Englewood Cliffs, NJ, 1989.
Suh, E., Lim, S., Hwang, H., and Kim, S. "A Prediction Model for the Purchase Probability of Anonymous Customers to Support Real Time Web Marketing: a Case Study," *Expert Systems with Applications* (27), 2004, pp. 245-255.
Wedel, M., Desarbo, W. S., Bult, J. R., and Ramaswamy, V. "A Latent Class Poisson Regression Model for Heterogeneous Count Data," *Journal of Applied Economics* (8), 1993, pp. 397 – 411.
Wei, C-P., and Chiu, I-T. "Turning Telecommunications Call Details to Churn Prediction: a Data Mining Ppproach," *Expert Systems with Applications* (23), 2002, pp.103-112.
Weng, S-S., and Liu, M-J. "Feature-Based Recommendation for One-to-one Marketing," *Expert Systems with Applications* (26), 2004, pp. 493-508.
Wirth, R., and Hipp, J. "CRISP-DM: Towards a Standard Process Model for Data Mining," *Proceedings of the 4th International Conference on the Practical Application of Knowledge Discovery and Data Mining*, Manchester, UK, 2000, pp. 29-39.
Yu, L., Liu, L., and Li, X. F. "A Hybrid Collaborative Filtering Method for Multiple-interests and Multiple-content Recommendation in E-commerce," *Expert Systems with Applications* (28:1), 2005, pp. 67-77.